Business Research Methods:
Data Analysis- I

By Dr. Satyabrata Dash

Professor- MBA Marketing

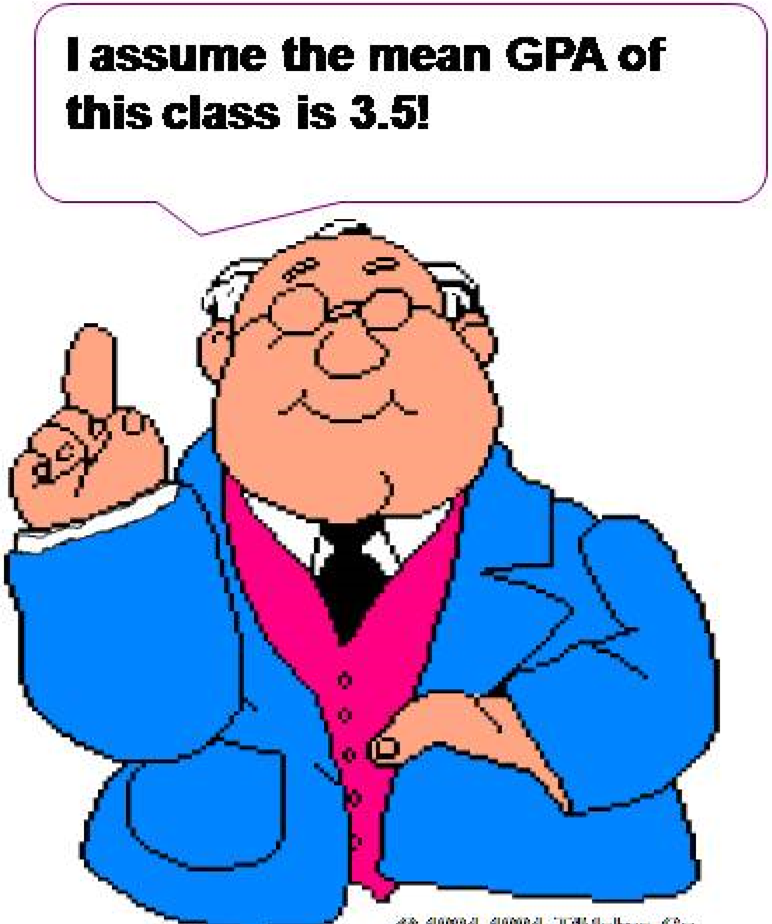
SMIT- PGCMS, Brahmapur

What is a Hypothesis?

- Hypothesis means proposition or supposition made as the basis for reasoning
- A hypothesis is an assumption about the population parameter.

A parameter is a characteristic of the population, like its mean or variance.

The parameter must be identified before analysis.



© 1984-1994 T/Maker Co.

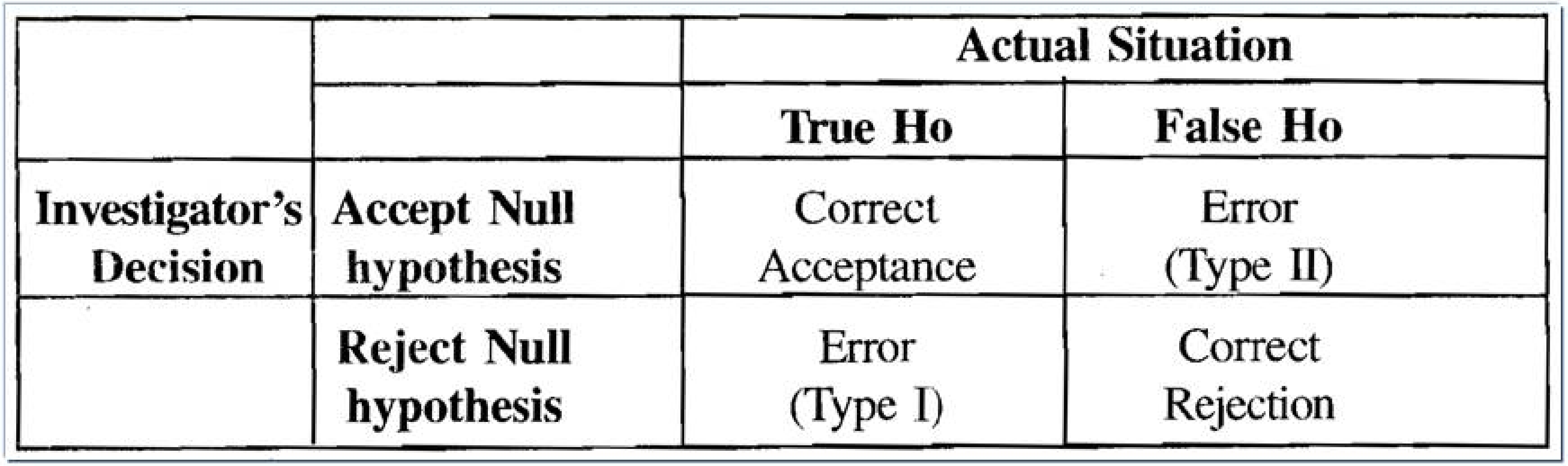
Classification of Hypothesis tests

- The researcher computes certain 'statistics' (sample values) as the basis for inferring the corresponding 'parameters' (population values).
- Ordinarily, a single sample is drawn from a given population so as to determine how well a researcher can infer or estimate the 'parameter' from a computed sample 'statistics'.
- For making the inferences about the various parameters, the researcher makes use of parametric and non-parametric tests.

Null and alternative hypotheses

- **Null hypothesis (H^0):** The null hypothesis is a claim of "no difference."
- **$H_0 : \mu = \mu_{H_0} = 100$**
- **Alternative hypothesis (H^a):** The alternative hypothesis is a claim of "a difference in the population," and is the hypothesis the researcher often hopes to bolster.
- **$H_a : \mu \neq \mu_{H_0} \neq 100$**
- It is important to keep in mind that the null and alternative hypotheses reference population values, and not observed statistics.

The concept of Type 1 and Type 2 Error



		Actual Situation	
		True Ho	**False Ho**
Investigator's Decision	**Accept Null hypothesis**	Correct Acceptance	Error (Type II)
	Reject Null hypothesis	Error (Type I)	Correct Rejection

p Value and conclusion

- Small *p values provide evidence against the null hypothesis because they say the observed data are unlikely when the null hypothesis is true. We apply the following* ***conventions:***
  - **If Calculated value < tabulated value then H_0 is accepted does not differ significantly**
  - **Calculated value > tabulated t value H_0 is rejected differ significantly**

PARAMETRIC TESTS

- Parametric tests are the most powerful statistical tests for testing the significance of the computed sampling statistics. These tests are based on the following assumptions:
  - the variables described are expressed in interval or ratio scales and not in nominal or ordinal scales of measurement,
  - the population values are normally distributed,
  - the samples have equal or nearly equal variances-this condition is known as 'equality or homogeneity of variances' and is particularly important to determine for small samples,
  - the selection of one case in the sample is not dependent upon the selection of any other.
- Application of parametric tests covers three tests, namely z-test, t-test and F-test.

Sampling Distribution of Means

- **A. Large Samples**
- An important principle, known as the 'central limit theorem', describes the characteristics of sample means. If a large number of equal-sized samples (greater than 30) are selected at random from an infinite population,
    - **the distribution of 'sample means' is normal and it possesses all the characteristics of a normal distribution,**
    - **the average value of 'sample means' will be the same as the mean of the population,**
    - **the distribution of the sample means around the population mean will have its own standard deviation, known as 'standard error of mean, which is denoted as SE_M or σ_M. It is computed by the formula**

$$\boxed{SE_M = \sigma_M = \frac{\bar{\sigma}}{\sqrt{N}}}$$

$$SE_M = \sigma_M = \frac{\bar{\sigma}}{\sqrt{N}} \quad \text{……………………(1)}$$

in which $\bar{\sigma}$ = Standard deviation of the population and
N = The number of cases in the sample.

Since the value of $\bar{\sigma}$ (i.e. standard deviation of population) is usually not known, we make an estimate of this standard error of mean by the formula:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad \text{…………………(2)}$$

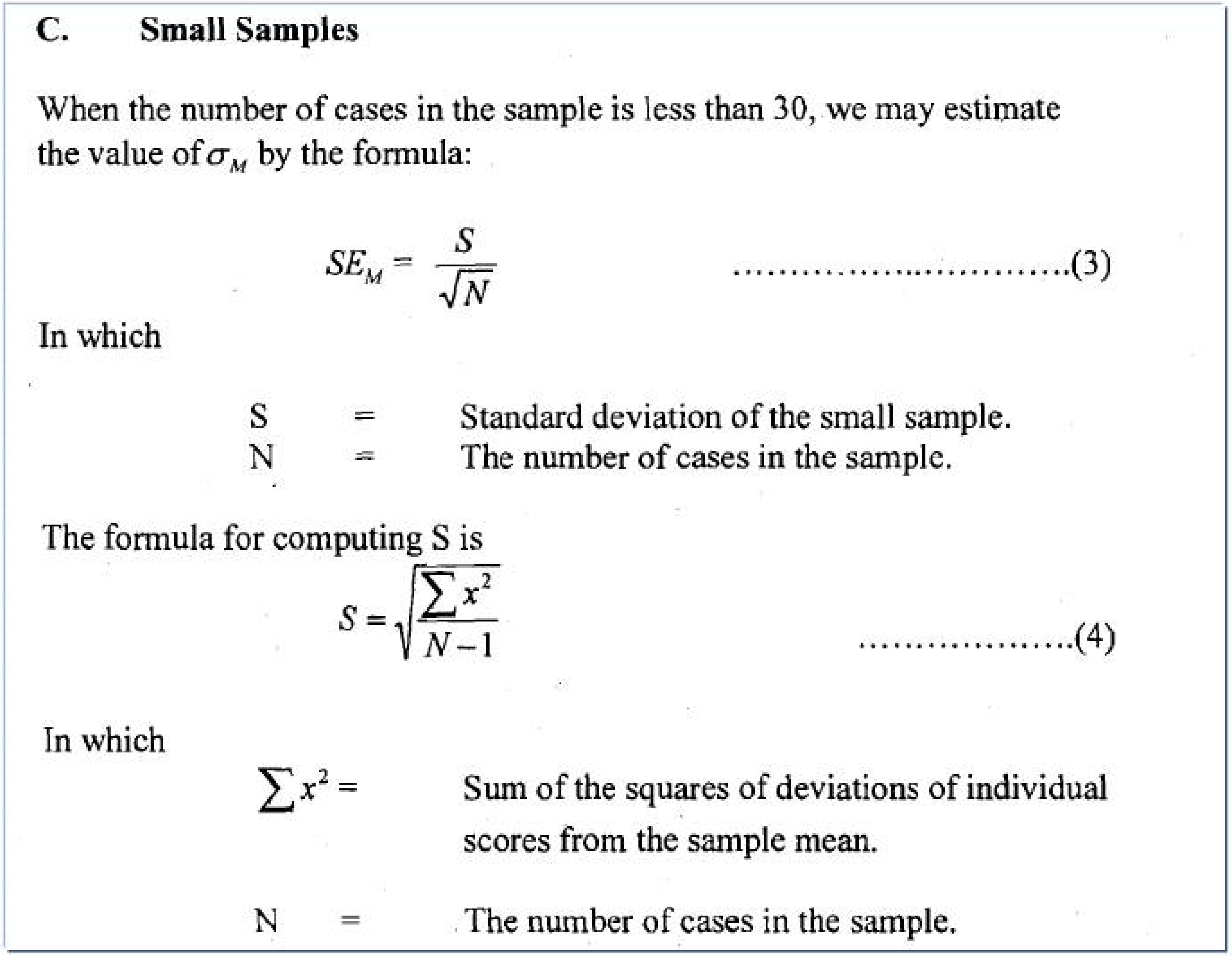
in which σ = Standard deviation of the sample
N = The number of cases in the sample.

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N-1}} = \sqrt{\frac{\sum x^2}{N-1}}$$

In which

$\sum x^2$ = Sum of the squares of deviations of individual scores from the sample mean.

N = The number of cases in the sample.



C. Small Samples

When the number of cases in the sample is less than 30, we may estimate the value of σ_M by the formula:

$$SE_M = \frac{S}{\sqrt{N}} \qquad \text{..............................(3)}$$

In which

S = Standard deviation of the small sample.
N = The number of cases in the sample.

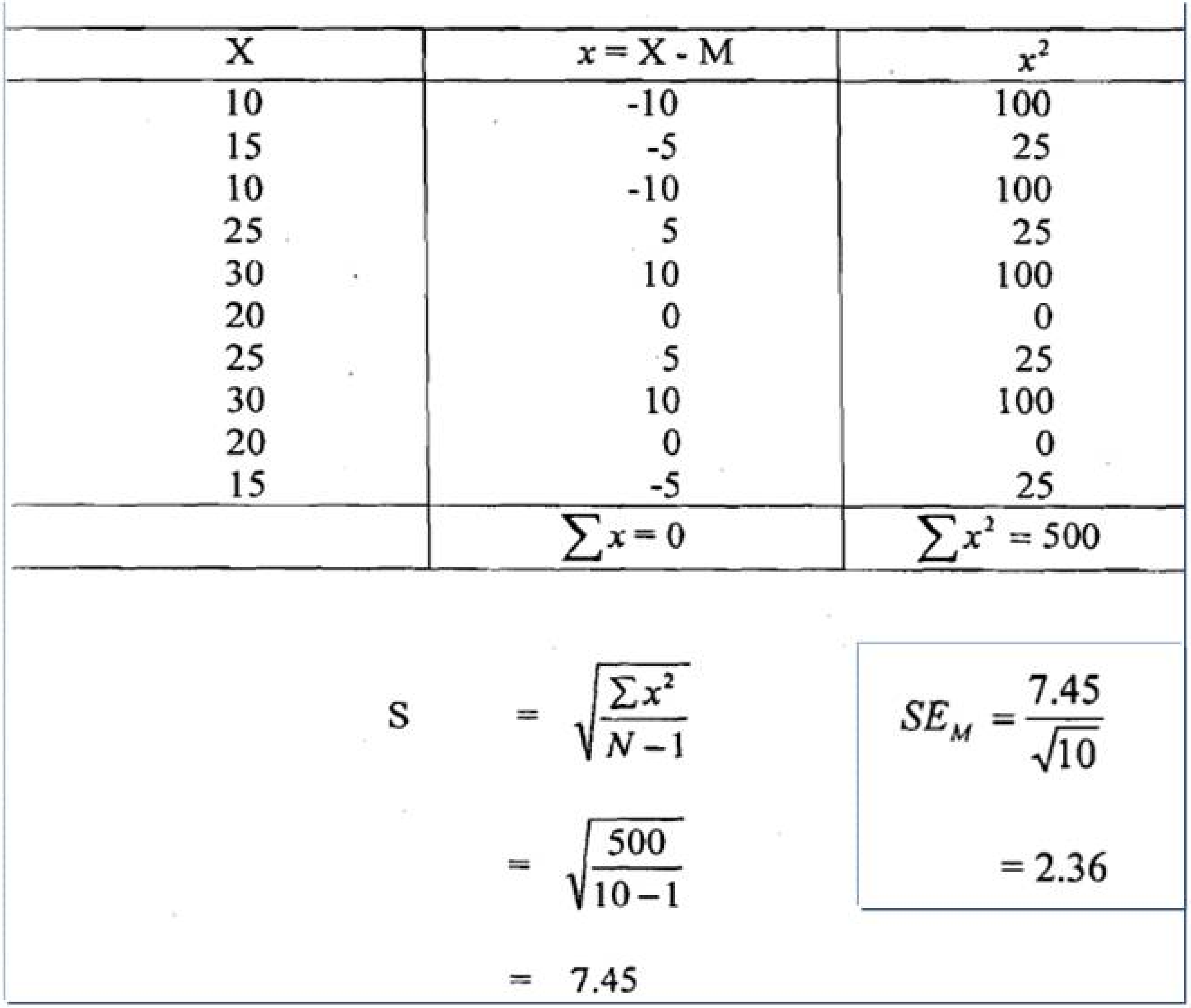
The formula for computing S is

$$S = \sqrt{\frac{\sum x^2}{N-1}} \qquad \text{..................(4)}$$

In which

$\sum x^2 =$ Sum of the squares of deviations of individual scores from the sample mean.

N = The number of cases in the sample.



X	$x = X - M$	x^2
10	-10	100
15	-5	25
10	-10	100
25	5	25
30	10	100
20	0	0
25	5	25
30	10	100
20	0	0
15	-5	25
	$\sum x = 0$	$\sum x^2 = 500$

$$S = \sqrt{\frac{\Sigma x^2}{N-1}}$$

$$= \sqrt{\frac{500}{10-1}}$$

$$= 7.45$$

$$SE_M = \frac{7.45}{\sqrt{10}}$$

$$= 2.36$$

The available df for determining t is N-1 or 9. we read that **SE$_M$** lies between table value 2.26 at 0.05 level and 3.25 at 0.01 level. Hence the probability shows 99% of confidence.

z-Test (Population is infinite)

n= 400 (Sample No.)

Mea
n X= 67.47 (Sample Mean)

μ_{H0}= 67.39 (Population Mean)

σ_p= 1.3 (S.D of Population)

N = ? (No. of Population)

Root of n	(Mean X- µH0)	σp/ Root of n
20	0.08	0.065

z= (Sample Mean - Population Mean)/ S.D of Population/ Root of n

z= (Mean X- µH0) / σp/ Root of n= 1.231

$H_0: \mu_{H0} = 67.39$

two tail test

$H_0: \mu_{H0} \neq 67.39$

Table value ot two tail z @5%level of significance = 1.96

Calculated t value < tabulated t value

H0 is accepted does not differ significantly

z-Test (Population is finite)

n= 5 (Sample No.)

Mean X= 300 (Sample Mean)

μ_{H0}= 320 (Population Mean)

σ_p= 75 (S.D of Population)

N = 20 (No. of Population)

Root of n	(Mean X- µH0)	(N-n)/N-1)	Root of ((N-n)/(N-1))	σp/ (Root of n)
2.24	-20	0.79	0.89	33.54

σp/ (Root of n)* Root of ((N-n)/(N-1))

29.80

z= (Sample Mean - Population Mean)/ S.D of Population/ Root of n

z= (Mean X- µH0) / σp/ (Root of n)* Root of ((N-n)/(N-1))= -0.67

H0: μ_{H0} = 67.39

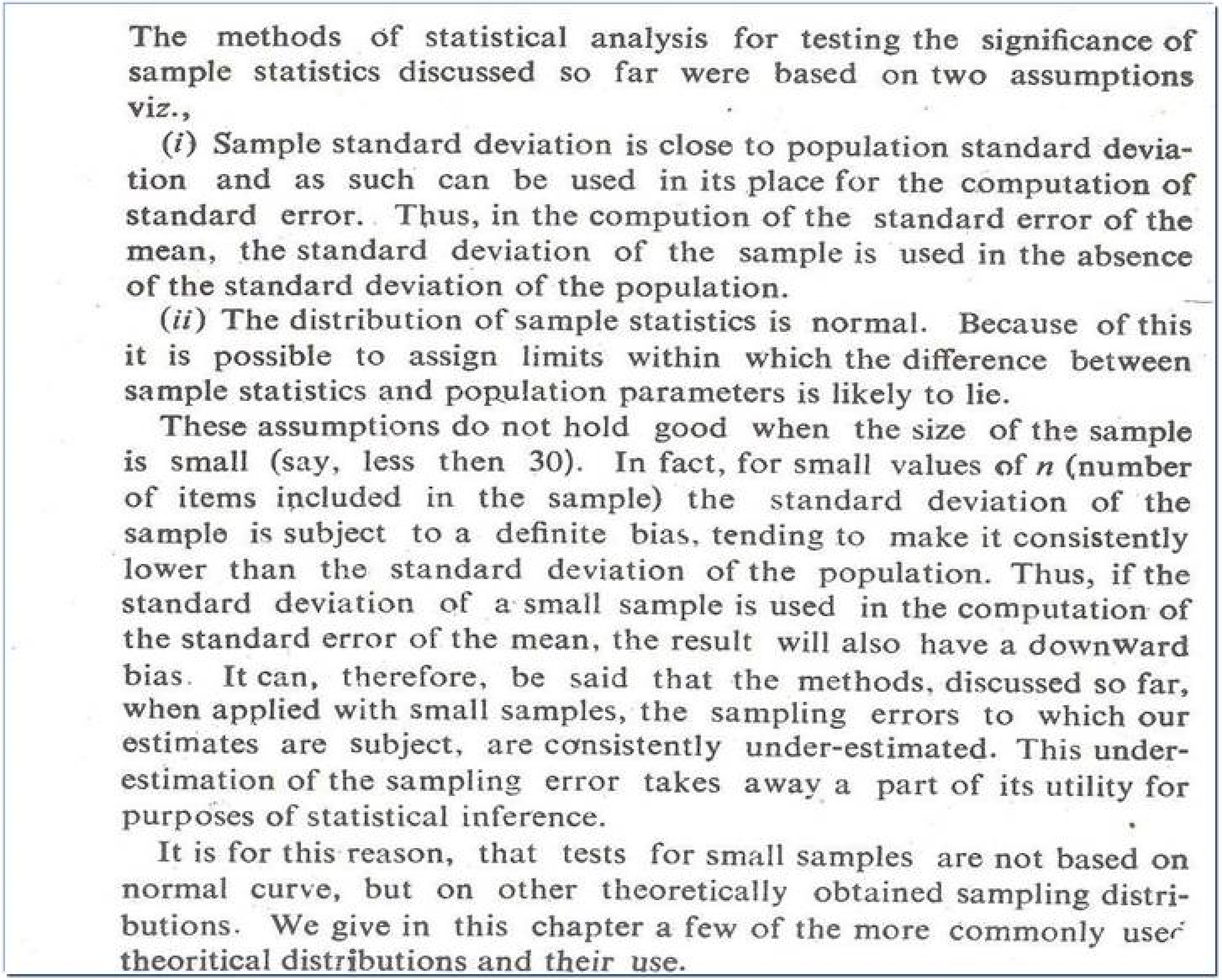
H0: $\mu_{H0} \neq$ 67.39 two tail test

Table value ot two tail z @5%level of significance = 1.96

Calculated t value < tabulated t value

H0 is accepted does not differ significantly

B. Small Samples



The methods of statistical analysis for testing the significance of sample statistics discussed so far were based on two assumptions viz.,

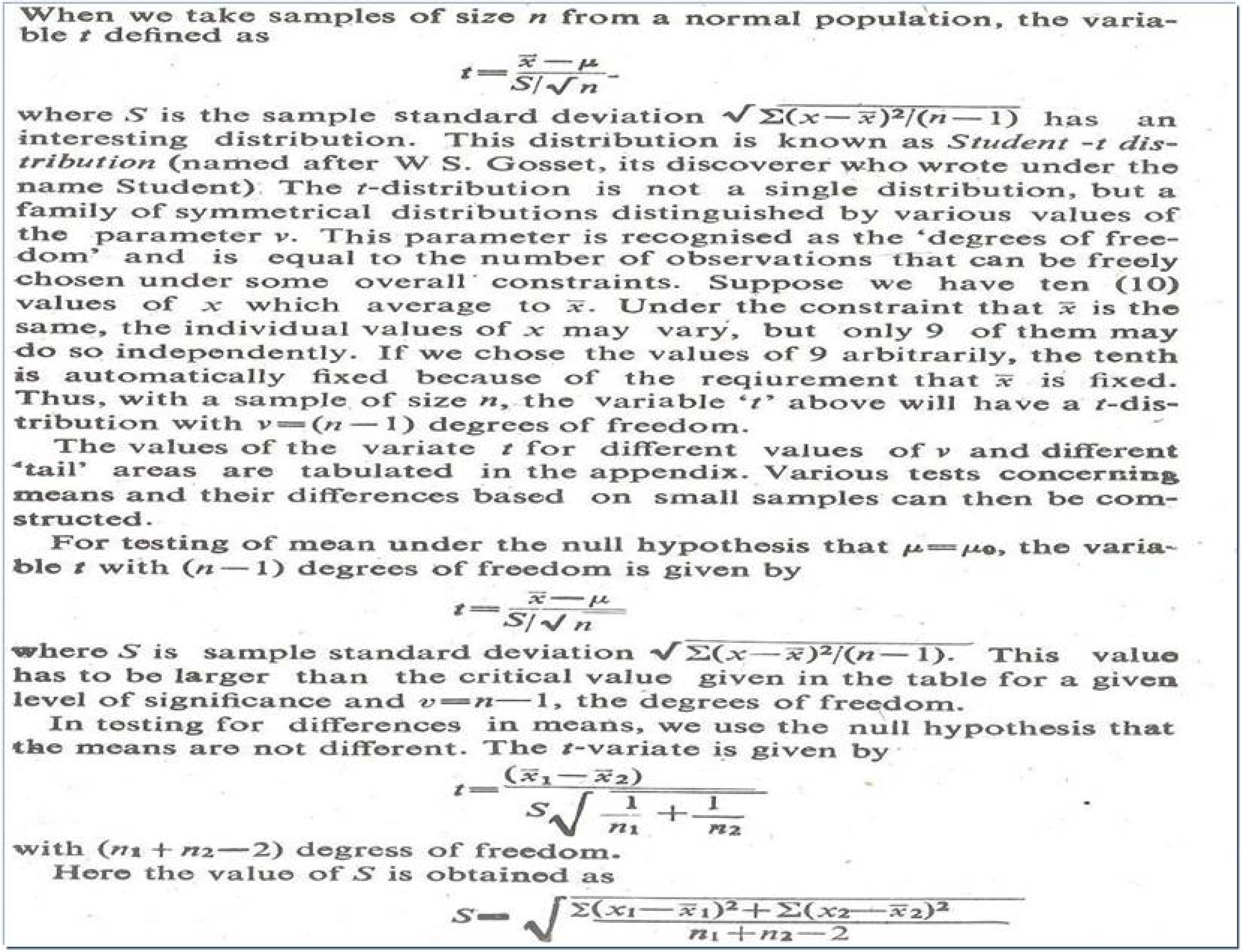
(*i*) Sample standard deviation is close to population standard deviation and as such can be used in its place for the computation of standard error. Thus, in the compution of the standard error of the mean, the standard deviation of the sample is used in the absence of the standard deviation of the population.

(*ii*) The distribution of sample statistics is normal. Because of this it is possible to assign limits within which the difference between sample statistics and population parameters is likely to lie.

These assumptions do not hold good when the size of the sample is small (say, less then 30). In fact, for small values of *n* (number of items included in the sample) the standard deviation of the sample is subject to a definite bias, tending to make it consistently lower than the standard deviation of the population. Thus, if the standard deviation of a small sample is used in the computation of the standard error of the mean, the result will also have a downward bias. It can, therefore, be said that the methods, discussed so far, when applied with small samples, the sampling errors to which our estimates are subject, are consistently under-estimated. This under-estimation of the sampling error takes away a part of its utility for purposes of statistical inference.

It is for this reason, that tests for small samples are not based on normal curve, but on other theoretically obtained sampling distributions. We give in this chapter a few of the more commonly used theoritical distributions and their use.

t-Test



When we take samples of size n from a normal population, the variable t defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

where S is the sample standard deviation $\sqrt{\Sigma(x-\bar{x})^2/(n-1)}$ has an interesting distribution. This distribution is known as *Student -t distribution* (named after W S. Gosset, its discoverer who wrote under the name Student). The t-distribution is not a single distribution, but a family of symmetrical distributions distinguished by various values of the parameter ν. This parameter is recognised as the 'degrees of freedom' and is equal to the number of observations that can be freely chosen under some overall constraints. Suppose we have ten (10) values of x which average to $\bar{x}$. Under the constraint that $\bar{x}$ is the same, the individual values of x may vary, but only 9 of them may do so independently. If we chose the values of 9 arbitrarily, the tenth is automatically fixed because of the reqiurement that $\bar{x}$ is fixed. Thus, with a sample of size n, the variable 't' above will have a t-distribution with $\nu=(n-1)$ degrees of freedom.

The values of the variate t for different values of ν and different 'tail' areas are tabulated in the appendix. Various tests concerning means and their differences based on small samples can then be constructed.

For testing of mean under the null hypothesis that $\mu=\mu_0$, the variable t with $(n-1)$ degrees of freedom is given by

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

where S is sample standard deviation $\sqrt{\Sigma(x-\bar{x})^2/(n-1)}$. This value has to be larger than the critical value given in the table for a given level of significance and $v=n-1$, the degrees of freedom.

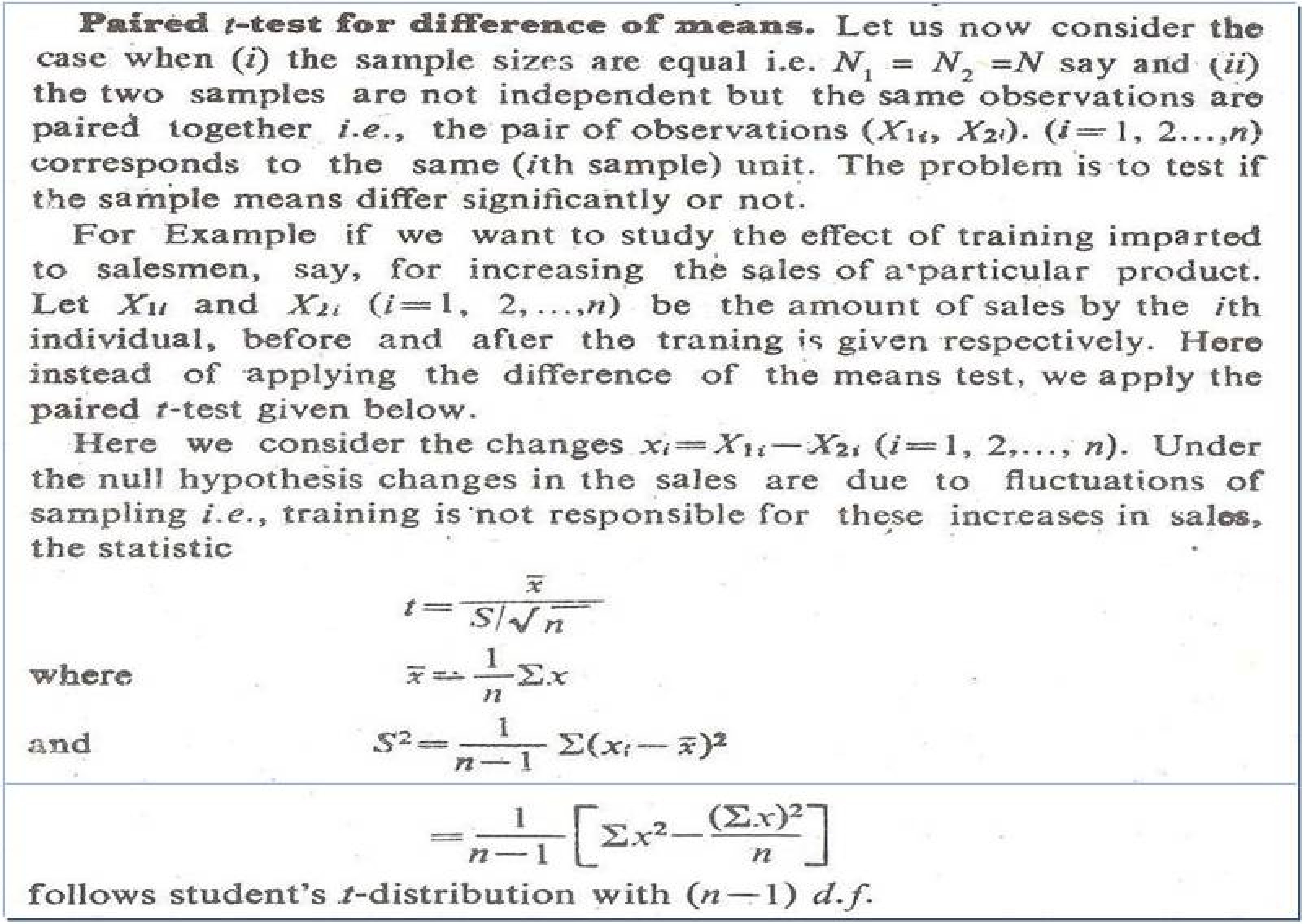
In testing for differences in means, we use the null hypothesis that the means are not different. The t-variate is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with (n_1+n_2-2) degress of freedom.

Here the value of S is obtained as

$$S = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}}$$



Paired t-test for difference of means. Let us now consider the case when (*i*) the sample sizes are equal i.e. $N_1 = N_2 = N$ say and (*ii*) the two samples are not independent but the same observations are paired together *i.e.*, the pair of observations (X_{1i}, X_{2i}). $(i = 1, 2 \ldots, n)$ corresponds to the same (*i*th sample) unit. The problem is to test if the sample means differ significantly or not.

For Example if we want to study the effect of training imparted to salesmen, say, for increasing the sales of a·particular product. Let X_{1i} and X_{2i} $(i=1, 2, \ldots, n)$ be the amount of sales by the *i*th individual, before and after the traning is given respectively. Here instead of applying the difference of the means test, we apply the paired t-test given below.

Here we consider the changes $x_i = X_{1i} - X_{2i}$ $(i = 1, 2, \ldots, n)$. Under the null hypothesis changes in the sales are due to fluctuations of sampling *i.e.*, training is not responsible for these increases in sales, the statistic

$$t = \frac{\bar{x}}{S/\sqrt{n}}$$

where $$\bar{x} = \frac{1}{n} \Sigma x$$

and $$S^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$$

$$= \frac{1}{n-1} \left[\Sigma x^2 - \frac{(\Sigma x)^2}{n} \right]$$

follows student's t-distribution with $(n-1)$ *d.f.*

One Sample t-Test

	Weight in Kg (X)	Deviation from Mean (x)= (X-Mean of X)	x^2
	110	-13	169
Mean weight of the population = 120	115	-8	64
	118	-5	25
	120	-3	9
	122	-1	1
	125	2	4
	128	5	25
	130	7	49
	139	16	256

$\sum X=$ 1107

$n=$ 9

Mean X = 123

$\sum x^2 =$ 602

σ= (Standard deviation of the sample) = Root over ($\sum x^2$/n-1)

($\sum x^2$/n-1)= 75.25

$\sigma =$ 8.67

$\mu =$ 120

Root of n = 3

t= 1.04 t= {(Mean X- μ) * Root of n}/σ

d.f= 8 d.f= (n- 1)

for 8 d.f @ 0.05 level of significance, the tabulated value of t = 2.31

Calculated t value < tabulated t value

H0 is accepted does not differ significantly

Two Samples t-Test

	Bacterium A	Bacterium B
Replicate 1	520	230
Replicate 2	460	270
Replicate 3	500	250
Replicate 4	470	280

$\sum x =$	1950	1030	(Total Sum of 4 Replicate value)
$n =$	4	4	
Mean $x =$	487.5	257.5	
$\sum x^2 =$	952900	266700	(Sum of the squares of each replicate value)
$(\sum x)^2 =$	3802500	1060900	Square of the total ($\sum x$). It is not the same as $\sum x$square
$(\sum x)^2/n =$	950625	265225	
$\sum d^2 =$	2275	1475	$\sum d^2 = \sum x^2 - (\sum x)^2/n$
$\sigma^2 =$	758.33	491.67	$\sigma^2 = \sum d^2/(n-1)$
		σd^2 is the variance of the difference $\sigma d^2 = \sigma 1^2/n1 + \sigma 2^2/n2$	
$\sigma d^2 =$	312.5		
$\sigma d =$	17.68		(the standard deviation of the difference between the means)
$t =$	13.01		$t = (\text{Mean } x1 - \text{Mean } x2)/\ \sigma d$
d.f=	6		d.f= $(n1 + n2) - 2$

for 6 d.f @ 0.05 level of significance, the tabulated value of t = 2.45

Calculated t value > tabulated t value

H0 is rejected   differ significantly

Difference of Means t-Test

	Before Change	After Change		Deviatn from mean	
Employee	(X1)	(X2)	X = (X2- X1)	(x)= (X- Mean of X)	x^2
A	24	26	2	1	1
B	26	26	0	-1	1
C	20	22	2	1	1
D	21	22	1	0	0
E	23	24	1	0	0
F	30	30	0	-1	1
G	32	32	0	-1	1
H	25	26	1	0	0
I	23	24	1	0	0
J	23	25	2	1	1

n= 10

Mean of X= 1

$\sum x^2 =$ 6

σ= (Standard deviation of the sample) = Root over ($\sum x^2$/n-1)

($\sum x^2$/n-1)= 0.67

σ = 0.82

Root of n = 3.16

t= 3.87 t= (Mean X * Root of n)/σ

d.f = 9 d.f= (n- 1)

for 9 d.f @ 0.05 level of significance, the tabulated value of t = 2.26

Calculated t value > tabulated t value

H0 is rejected differ significantly

Analysis of Variance (ANOVA) or F Test

- Want to study the effect of one or more ***qualitative variables on a quantitative*** outcome variable
- Qualitative variables are referred to as ***factors***
- Characteristics that differentiates factors are referred to as ***levels (i.e., three genotypes of a*** SNP

One and Two Sided Tests

- Hypothesis tests can be one or two sided (tailed)

- One tailed tests are directional:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_A: \mu_1 - \mu_2 > 0$$

- Two tailed tests are not directional:

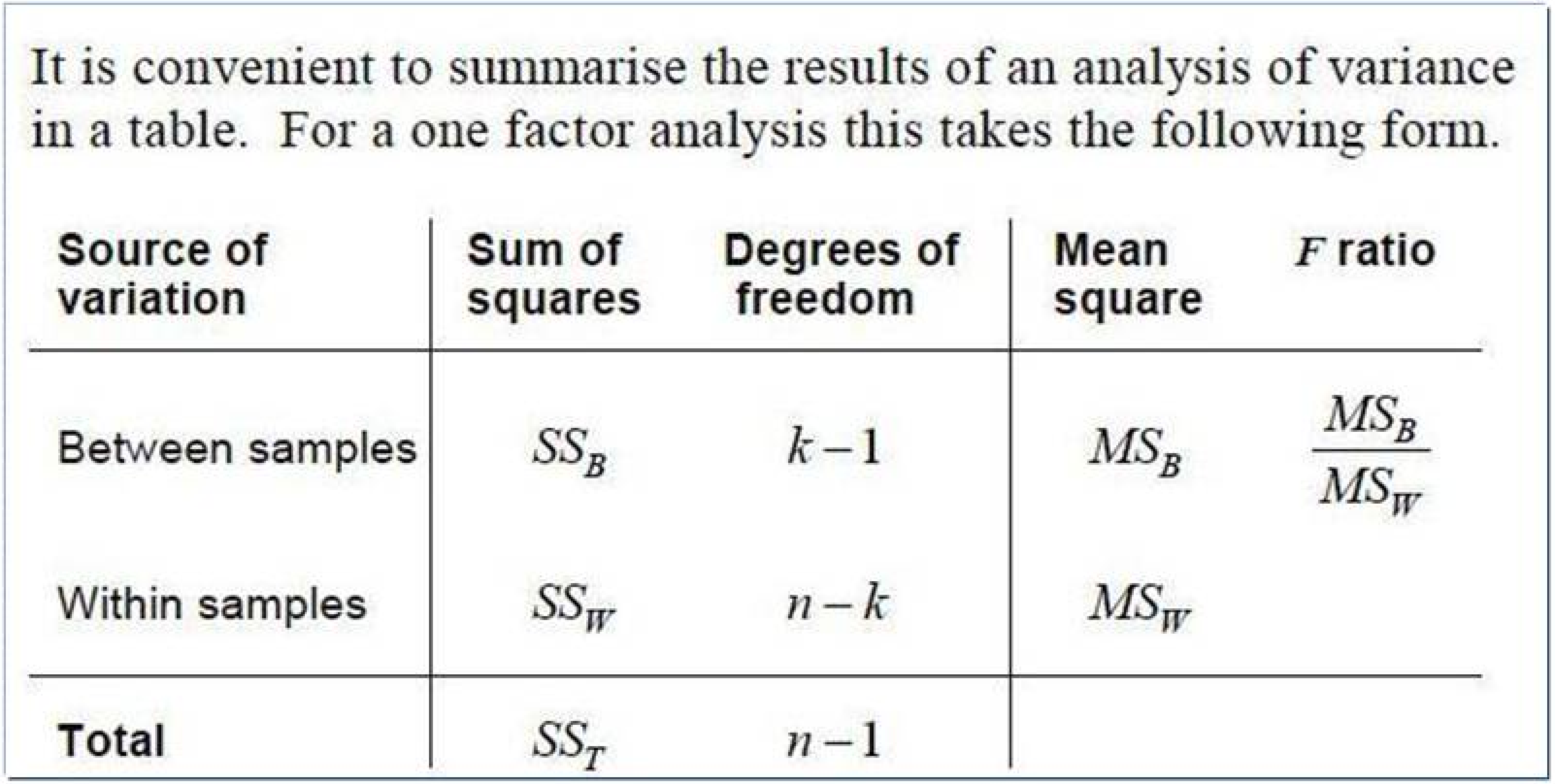
$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

One-Way ANOVA

- **Simplest case is for One-Way (Single Factor) ANOVA**
  - The *outcome variable is the variable you're comparing*
  - The *factor variable is the categorical variable being used to* define the groups
    - We will assume *k samples (groups)*
  - The *one-way is because each value is classified in exactly one* way
- **ANOVA easily generalizes to more factors**

One way ANOVA Table



It is convenient to summarise the results of an analysis of variance in a table. For a one factor analysis this takes the following form.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between samples	SS_B	$k-1$	MS_B	$\dfrac{MS_B}{MS_W}$
Within samples	SS_W	$n-k$	MS_W	
Total	SS_T	$n-1$		

One way ANOVA test

X1	X2	X3
1	2	2
2	4	3
5	2	4

Step- 2

df between = k-1 = 3-1 = 2

where k= no. of groups = 3

df within = N-k = 9-3 = 6

Where N= Total no. of scores = 9

df Total = (N-1) = 8

F critical (refer table df between : numerator & df within : denominator) 5.14

Step-1

H0= μ1 = μ2 = μ3

Ha = At least 1 difference among the means

ά = 0.05

Step- 3

Mean X1 =	2.67
Mean X2 =	2.67
Mean X3 =	3
G/N =	2.78

Sum of Square between (SS between) = SS Total- SS within

(SS between)
0.22

Sum of Square Total (SS Total)= Sum Square (Xi- G)

Square (Xi- G)			SS Total
3.16	0.60	0.60	13.56
0.60	1.49	0.05	
4.94	0.60	1.49	

Sum of Square Within (SS within) = Sum Square (Xi- MeanXi)

Square (Xi- Mean Xi)			SS within
2.78	0.44	1.00	13.33
0.44	1.78	0.00	
5.44	0.44	1	

Step- 4 Calculate Variance

Mean Square between (MS between)= SS between/df between= **0.11**

Mean Square within (MS within)= SS within/df within = **2.22**

Step- 5

F= (MS between)/ (MS within)= **0.05**

F critical value = (Table Value)= **5.14**

0.05 < 5.14 Fail to Reject H0

or H0= µ1 = µ2 = µ3

There is no significant difference between three groups

One way ANOVA test

X1	X2	X3
82	83	38
83	78	59
97	68	55

Step-1

H0= µ1 = µ2 = µ3

Ha = Atleast 1 difference among the means

ά = 0.05

Step- 2

df between = k-1 = 3-1 = 2

where k= no. of groups = 3

df within = N-k = 9-3 = 6

Where N= Total no. of scores = 9

df Total = 8

F critical (refer table df betw = numerator & df witn denominator) 5.14

Step- 3

Mean X1 =	87.3
Mean X2 =	76.3
Mean X3 =	50.7
G/N =	71.4

Sum of Square Total (SS Total)= Sum Square (Xi- G)

Sum of Square Within (SS within) = Sum Square (Xi- MeanXi)

Sum of Square between (SS between) = SS Total- SS within

Mean Square between (MS between)= SS between/df between

F= (MS between)/ (MS within)

Sum Square (Xi- G)			SS Total	Sum Square (Xi- Mean Xi)			SS within	(SS between)
111.42	133.53	1118.53	**2630.22**	28.44	44.44	160.44	**506.00**	**2124.22**
133.53	42.98	154.86		18.78	2.78	69.44		
653.09	11.86	270.42		93.44	69.44	18.78		

Step- 4 Calculate Variance

Mean Square between (MS between)= SS between/df between= **1062.11**

Mean Square within (MS within)= SS within/df within = **84.33**

Step- 5

F= (MS between)/ (MS within)= **12.59**

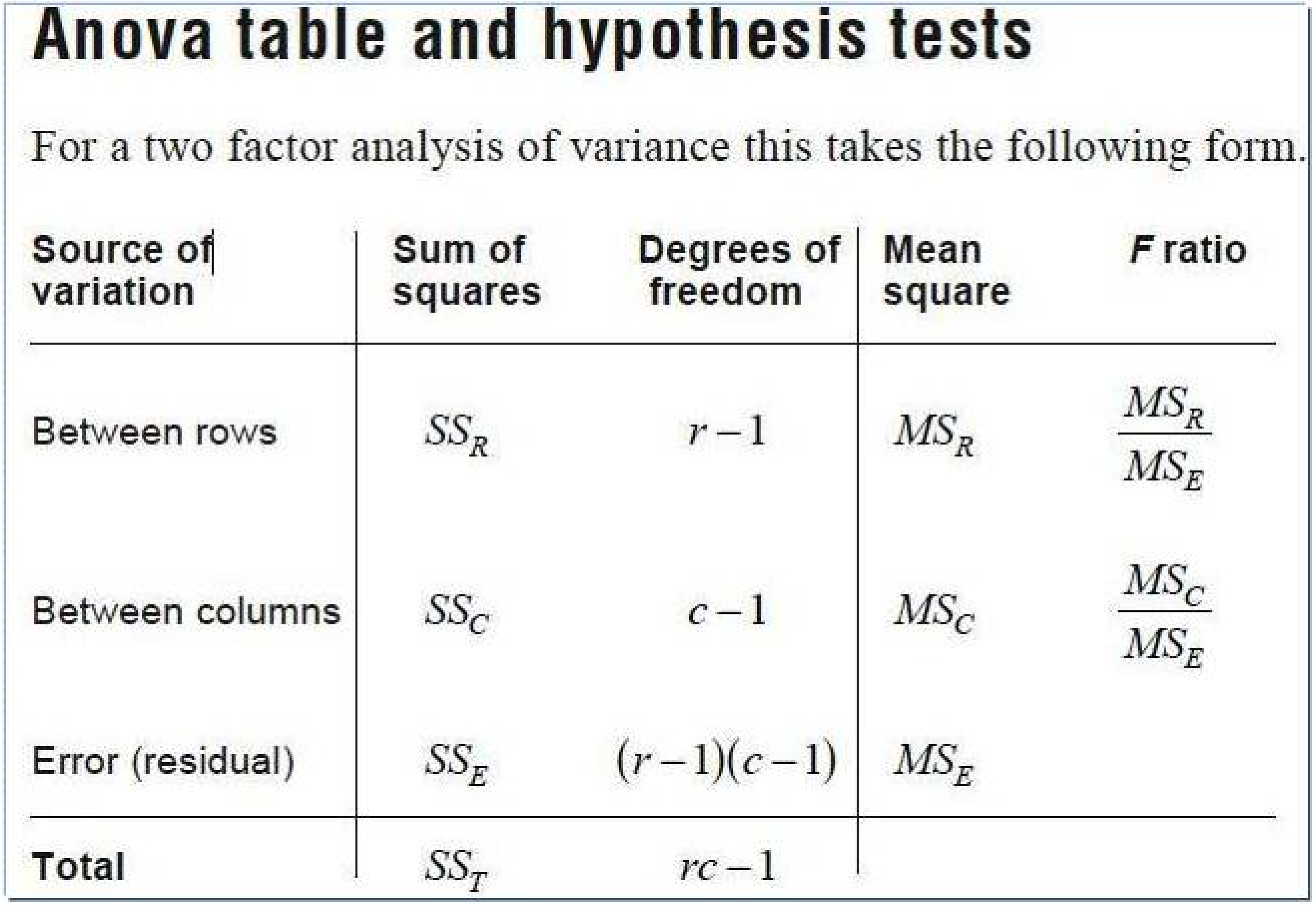
F critical value = (Table Value)= **5.14**

12.59 > 5.14 Reject H0

or H0≠ µ1 ≠ µ2 ≠ µ3

There is significant difference between three groups

Two way ANOVA Table



Anova table and hypothesis tests

For a two factor analysis of variance this takes the following form.

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* ratio
Between rows	SS_R	$r-1$	MS_R	$\frac{MS_R}{MS_E}$
Between columns	SS_C	$c-1$	MS_C	$\frac{MS_C}{MS_E}$
Error (residual)	SS_E	$(r-1)(c-1)$	MS_E	
Total	SS_T	$rc-1$		

Two way ANOVA test

Fertilizers/ Seeds	a	b	c
w	6	5	5
x	7	5	4
y	3	3	3
z	8	7	4

Step-1

T= 60 (Sum of all)

n= 12 Row * Coulmn

Correction Factor= Square of T/n 300

Step- 2

Total SS= Total Sum of Square - square of T/n

Total Square

36	25	25
49	25	16
9	9	9
64	49	16

Sum of Square 332

Total SS= 332-300 32

Fertilizers/Seeds	a	b	c	SR	SSR = $(SR)^2/(ni-1)$
w	6	5	5	16	85.33
x	7	5	4	16	85.33
y	3	3	3	9	27.00
z	8	7	4	19	120.33
SC	24	20	16		
SSC= $(SC)^2/(ni-1)$	144	100	64		

Step-3

SS between column treatment= Square of Sum of column items- square of T/n

SS between column treatment= **8**

Step-4

SS between row treatment= Square of Sum of row items- square of T/n

SS between row treatment= **18.00**

Step- 5

SS residual or

Error= Total SS- (SS between column + SS between row)

6.00

The ANOVA Table

Source of variation	SS	df	MS	F-ratio	5% F limit or table value		
Between column	8	2	4	4	F(2,6)= 5.14	not significant	Accept H0
Between row	18	3	6	6	F(3,6)= 4.76	Significant	Reject H0
Residual error	6	6	1				
Total	32	11					

df between column= (No of column-1)

df between row= (No of row-1)

df residual error= (No of column-1)* (No of row-1)

MS= SS/df

F-ratio= MS/Residual error at MS

NB

if the calculated f value > table value= Significant

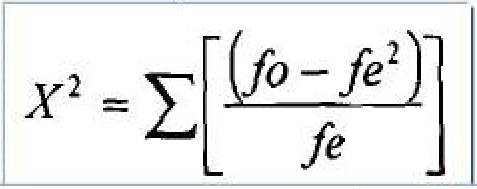
if the calculated f value< table value= not significant

Non-Parametric Test

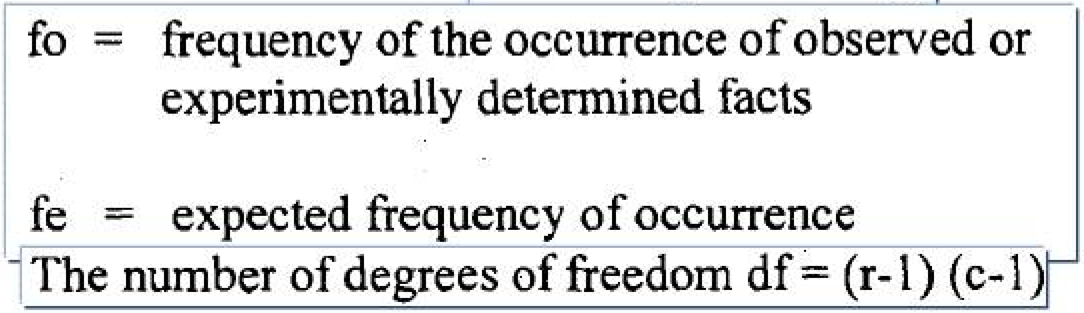
- Many of the hypothesis tests require normal distributed populations or some tests require that population variances be equal. What if, for a given test, such requirements cannot be met? For these cases, statisticians have developed hypothesis tests that are "distribution free." Such tests are called nonparametric tests.
- A nonparametric test is a hypothesis test that does not require any specific conditions concerning the shape of populations or the value of any population parameters.
- Nonparametric tests are easier to perform (they do not require normally distributed populations).
- They can be applied to categorical data (such as genders of survey responds).
- They are less efficient than parametric tests. Stronger evidence is required to reject a null hypothesis.

Chi-Square (χ^2) Test

- The Chi-square (pronounced as Ki-square) test is used with discrete data in the form of frequencies. It is a test of independence and is used to estimate the likelihood that some factor other than chance accounts for the observed relationship. Since the null hypothesis states that there is no relationship between the variables under study, the Chi-square test merely evaluates the probability that the observed relationship results from chance. The formula for Chi-square is



$$X^2 = \sum\left[\frac{(fo - fe^2)}{fe}\right]$$



fo = frequency of the occurrence of observed or experimentally determined facts

fe = expected frequency of occurrence

The number of degrees of freedom df = (r-1) (c-1)

Chi-Square (χ^2) Test

	Favor	Neutral	Oppose	f row
Democrat	10	10	30	50
Republican	15	15	10	40
f column	25	25	40	90

Row Frequency = (f row)

Column Frequency = (f column)

n= 90

Level of Significance (ά) = 0.05

H0 = There is no Significant difference between

Ha = There is association between

d.f. = (R-1)(C-1) = 2

Calculate Expected friquency = $f_e = f_r f_c / n$

	Favor	Neutral	Oppose
Democrat	13.89	13.89	22.22
Republican	11.11	11.11	17.78

$$\chi^2 = \sum \left[\frac{(F_o - F_e)^2}{F_e} \right]$$

$\boxed{\chi^2}$ = 11.03

Critical tabled value of α = 0.05 at d.f of 2 = 5.991

Calculated t value > tabulated t value

H0 is rejected differ significantly

fo - fe

-3.89	-3.89	7.78
3.89	3.89	-7.78

$(fo - fe)^2$

15.12	15.12	60.49
15.12	15.12	60.49

$(fo - fe)^2/fe$

1.09	1.09	2.72
1.36	1.36	3.40

The Kruskal- Wallis Test

- The Kruskal- Wallis test is the version of the independent measures (one-way) ANOVA that can be performed on ordinal (ranked) data.
- The only requirement for Kruskal- Wallis test are:

1- The k sample are random and independent.

2- There are 5 or more measurements per sample.

3- The probability distributions are conteneous.

Sample-1	Sample-2	Sample-3
8.2	10.2	13,5
10.3	9.1	8.4
9.1	13.9	9.6
12.6	14.5	13.8
11.4	9.1	17.4
13.2	16.4	15.3

H0: the 3 probability distributions are identical

or

Ha: At least 2 of the 3 probability distributions differ in location.

Step-2

Sample-1	Rank	Sample-2	Rank	Sample-3	Rank
8.2	1	10.2	7	13,5	12
10.3	8	9.1	4	8.4	2
9.1	4	13.9	14	9.6	6
12.6	10	14.5	15	13.8	13
11.4	9	9.1	4	17.4	18
13.2	11	16.4	17	15.3	16
	43		61		67

R1 = 43 R2 = 61 R3 = 67

n1 = 6 n2 = 6 n3 = 6

K = 3 n = 18

Step-1

Sample Scores	Ranking	Ranking
(Order)	(Rough)	
8.2	1	1
8.4	2	2
9.1	3	4
9.1	4	4
9.1	5	4
9.6	6	6
10.2	7	7
10.3	8	8
11.4	9	9
14.5	10	10
12.6	11	11
13.2	12	12
13,5	13	13
13.8	14	14
13.9	15	15
15.3	16	16
16.4	17	17
17.4	18	18

$$H = 12/\ n(n+1)\sum_{i=1}^{k} R_i^2/\ n_i\ -\ 3(n+1)$$

$12/n(n+1) =$ 0.035

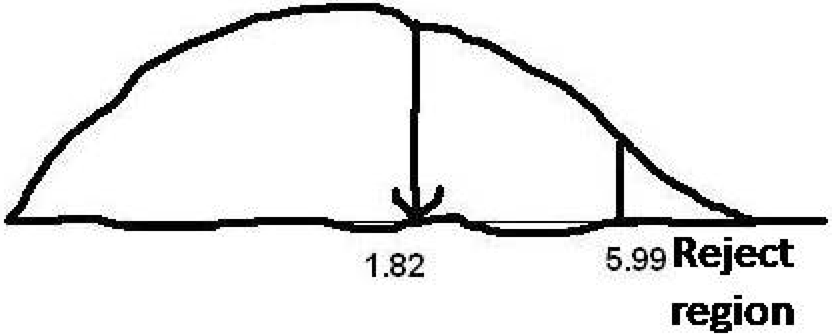
$\sum R_i^2/n_i =$ $R1^2/n1 + R2^2/n2 + R3^2/n3$ 1676.5

$3(n+1) =$ 57

H = 1.8245614 (Test Statistics)

Step- 3

- Rejection region (Chi Square)
- RR: $H > X^2_{ά,\ k-1}$ $ά = 0.05$ d.f= k-1= 2
- RR: H> 5.99



Step- 4 Decision

- Rejection region (Chi Square)
- RR: H> X
- RR: H> 5.99

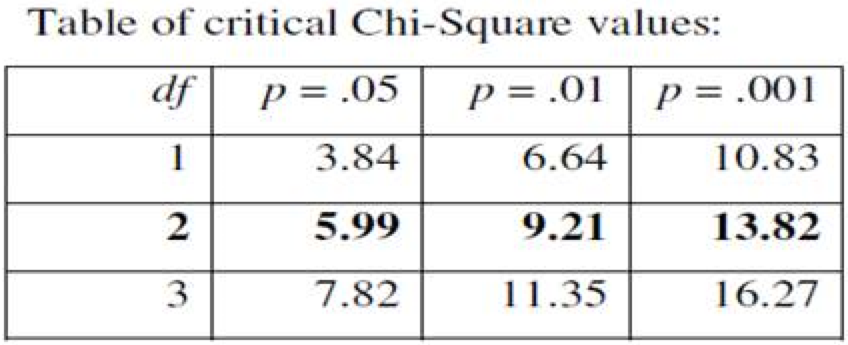
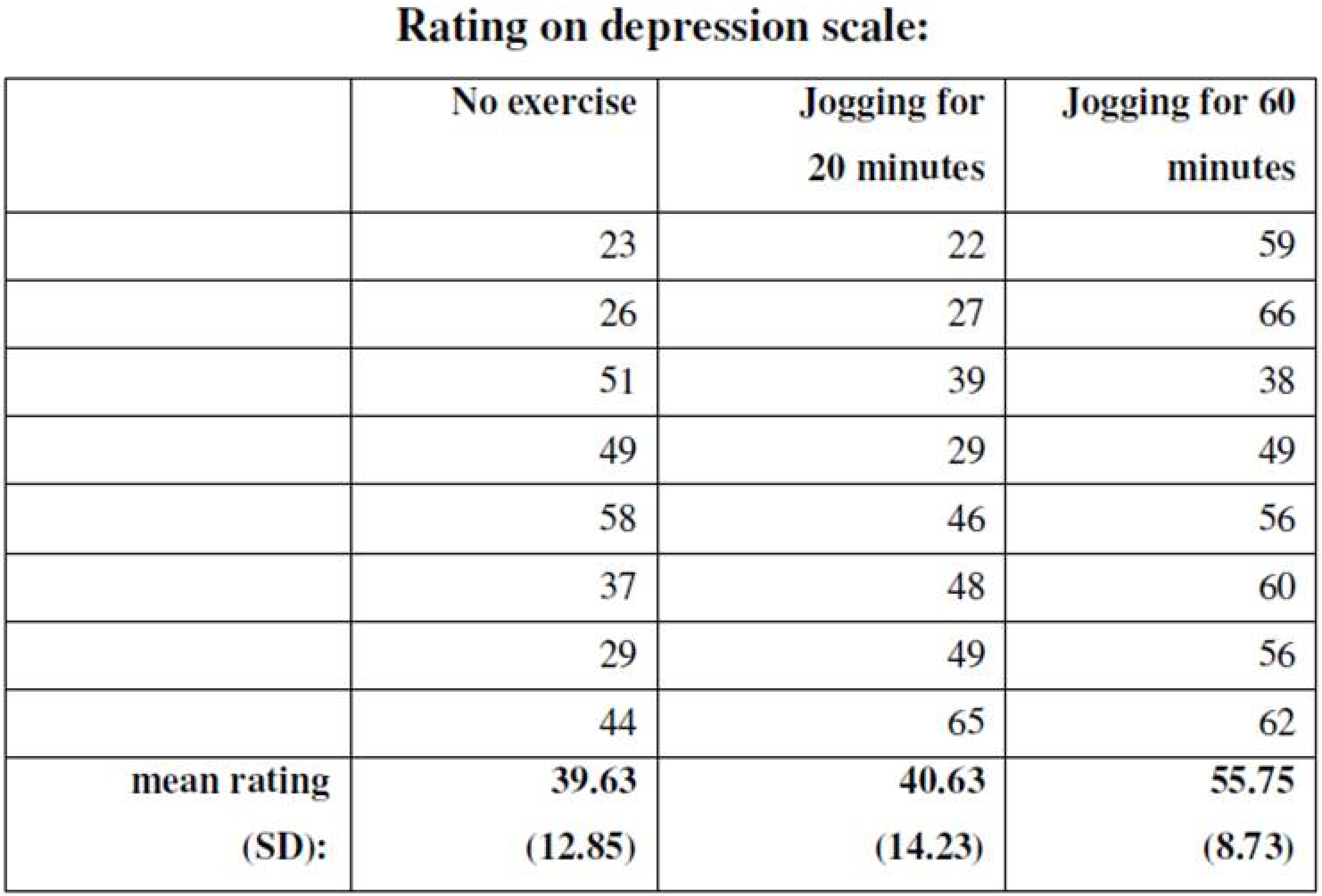


Table of critical Chi-Square values:

df	*p* = .05	*p* = .01	*p* = .001
1	3.84	6.64	10.83
2	**5.99**	**9.21**	**13.82**
3	7.82	11.35	16.27

- **Step- 5 Conclusion:** Do not reject H_0

Example



Rating on depression scale:

	No exercise	Jogging for 20 minutes	Jogging for 60 minutes
	23	22	59
	26	27	66
	51	39	38
	49	29	49
	58	46	56
	37	48	60
	29	49	56
	44	65	62
mean rating **(SD):**	**39.63** **(12.85)**	**40.63** **(14.23)**	**55.75** **(8.73)**

H0: the 3 probability distributions are identical

or

Ha: At least 2 of the 3 probability distributions differ in location.

Step- 2

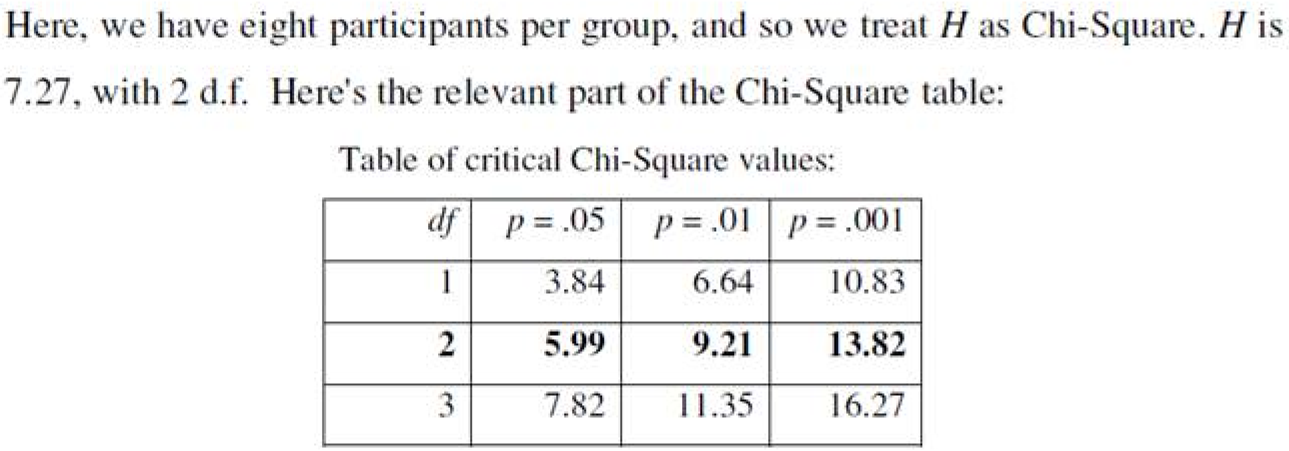
Sample- 1	Rank	Sample- 2	Rank	Sample- 3	Rank
23	2	22	1	59	20
26	3	27	4	66	24
51	16	39	9	38	8
49	14	29	5.5	49	14
58	19	46	11	56	17.5
37	7	48	12	60	21
29	5.5	49	14	56	17.5
44	10	65	23	62	22
	76.5		79.5		144

R1 =77 R2 = 80 R3 = 144

n1= 8 n2 = 8 n3= 8

K= 3 n= 24

Sample Scores (Order)	Ranking
22	1
23	2
26	3
27	4
29	5.5
29	5.5
37	7
38	8
39	9
44	10
46	11
48	12
49	14
49	14
49	14
51	16
56	17.5
56	17.5
58	19
59	20
60	21
62	22
65	23
66	24



Here, we have eight participants per group, and so we treat *H* as Chi-Square. *H* is 7.27, with 2 d.f. Here's the relevant part of the Chi-Square table:

Table of critical Chi-Square values:

df	*p* = .05	*p* = .01	*p* = .001
1	3.84	6.64	10.83
2	**5.99**	**9.21**	**13.82**
3	7.82	11.35	16.27

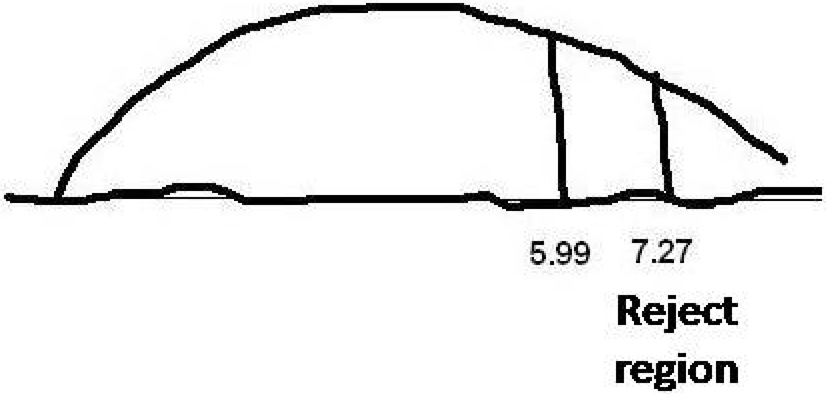
- Rejection region (Chi Square)
- RR: H> $X^2_{ά, k-1}$
- RR: H> 5.99

ά = 0.05

d.f= k-1= 2

Decision

- Rejection region (Chi Square)
- RR: H> X
- RR: H> 5.99



Do reject H_0

One-Sample Sign Test

Daily emission of SO2

X	µ(given)	X-µ	(+/-)
17	23.5	-7	(-)
15	23.5	-9	(-)
20	23.5	-4	(-)
29	23.5	6	(+)
19	23.5	-5	(-)
18	23.5	-6	(-)
22	23.5	-2	(-)
25	23.5	2	(+)
27	23.5	4	(+)
9	23.5	-15	(-)
24	23.5	1	(+)
20	23.5	-4	(-)
17	23.5	-7	(-)
6	23.5	-18	(-)
24	23.5	1	(+)
14	23.5	-10	(-)
15	23.5	-9	(-)
23	23.5	-1	(-)
24	23.5	1	(+)
26	23.5	3	(+)
19	23.5	-5	(-)
23	23.5	-1	(-)
28	23.5	5	(+)
19	23.5	-5	(-)
16	23.5	-8	(-)
22	23.5	-2	(-)
24	23.5	1	(+)
17	23.5	-7	(-)
20	23.5	-4	(-)
13	23.5	-11	(-)
19	23.5	-5	(-)
10	23.5	-14	(-)
23	23.5	-1	(-)
18	23.5	-6	(-)
31	23.5	8	(+)
13	23.5	-11	(-)
20	23.5	-4	(-)
17	23.5	-7	(-)
24	23.5	1	(+)
14	23.5	-10	(-)

Success	Failure
11	29

Ho: p = 1/2

Ha: p< 1/2

X= 11

n= 40

P= 1/2

Z= (x-npo)/Root of ((npo*(1-po))

x- npo = -9

npo*(1-po)= 10

Root of ((npo*(1-po))= 3.16

Z= -2.85

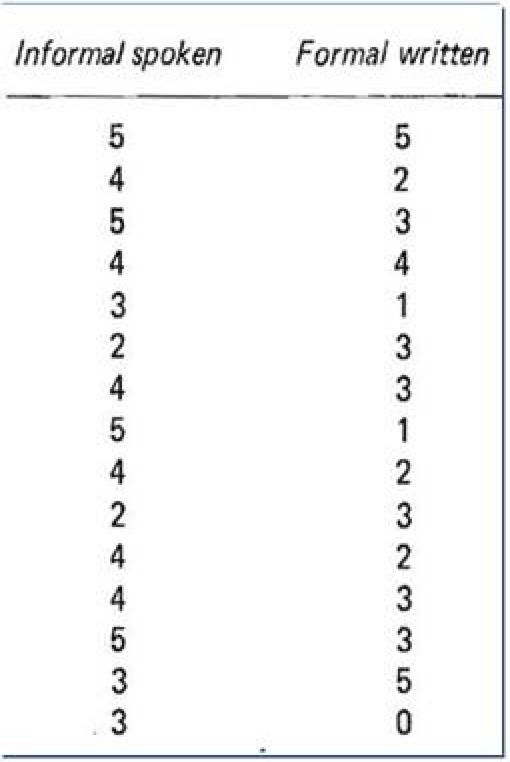
Absolute Z = !Z!= 2.85

Tabulated value of Z at ά = 0.05= 1.645

Calculated value > tabulated value

H0 is rejected differ significantly

Two-Sample Sign Test



Informal spoken	*Formal written*
5	5
4	2
5	3
4	4
3	1
2	3
4	3
5	1
4	2
2	3
4	2
4	3
5	3
3	5
3	0

X	Y	X-Y	(+/-)
5	5	0	0
4	2	2	(+)
5	3	2	(+)
4	4	0	0
3	1	2	(+)
2	3	-1	(-)
4	3	1	(+)
5	1	4	(+)
4	2	2	(+)
2	3	-1	(-)
4	2	2	(+)
4	3	1	(+)
5	3	2	(+)
3	5	-2	(-)
3	0	3	(+)

Success	Failure
10	3

Z= (x-npo)/Root of ((npo*(1-po))

x- npo = 3.5

npo*(1-po)= 3.25

Root of ((npo*(1-po))= 1.80

Ho: p = 1/2

Ha: p> 1/2

X= 10

n= 13

P= 1/2

Z= 1.94

Absolute Z = !Z!= 1.94

Tabulated value of Z at ά = 0.05= 1.645

Calculated value > tabulated value

H0 is rejected differ significantly

Runs Test for Randomness

In order to draw conclusions about the population on the basis of the sample information, it is necessary that the sample drawn must be random or unbiased. The runs test is used to test the sample for randomness. The test is based on the order or sequence in which the individual observations originally were obtained. A run is defined as a sequence of identical symbols or elements which are followed and proceeded by different types of symbols or elements or by no symbols on either side.

For example, in studying the arrival pattern of customers in a large departmental store, we might observe the following sequence of male (M) and female (F) arrivals

M M F F F M M F F F M M M M F M M F F M

Runs Test

Ho= The arrival pattern , sex wise, of the customers at the super market is random
Ha= The arrival pattern , sex wise, of the customers at the super market is not random

Sex-wise arrival pattern in Super Market

MM
WWW
M
WW
MM
WWWW
MMM
WW
MM
W
MMM
WWW
MM
WW
MM
WW
M
WW
MM
WW
M
WW
MM
WW

n1 = Man = 23
n2 = Woman = 27
r = pattern = 24

2n1n2 =	1242
2n1n2 (2n1n2 -n1- n2) =	1480464
$(n1+n2)^2$ =	2500
n1+n2-1 =	49
$SD (r)^2$ =	12.09

$$\text{Mean}= E(r)= \frac{2n_1\ n_2}{n_1 + n_2} + 1 = 25.84$$

$$SD (r) = \sqrt{\frac{2n_1\ n_2\ (2n_1\ n_2 - n_1 - n_2)}{(n_1 + n_2)^2\ (n_1 + n_2 - 1)}} = 3.48$$

$$|z| = \frac{|r - E(r)|}{S.D.(r)} = 0.52928 \quad 0.53$$

Tabulated value of Z at ά = 0.05= 1.96

Calculated Z value < tabulated Z value

H0 is accepted

Hence the arrival pattern , sex wise, of the customers at the super market is random